UNLIMITED
UNCLASSIFIED

Canada

AN OVERVIEW OF TECHNOLOGY FOR SPOKEN INTERACTION WITH MACHINES

by
M.J. Hunt
National Aeronautical Establishment

OTTAWA
FEBRUARY 1988

AERONAUTICAL NOTE
NAE-AN-50
NRC NO. 28714

National Research Council Canada
Conseil national de recherches Canada

88-03209/

NATIONAL AERONAUTICAL ESTABLISHMENT

SCIENTIFIC AND TECHNICAL PUBLICATIONS

AERONAUTICAL REPORTS:

Aeronautical Reports (LR): Scientific and technical information pertaining to aeronautics considered important, complete, and a lasting contribution to existing knowledge.

Mechanical Engineering Reports (MS): Scientific and technical information pertaining to investigations outside aeronautics considered important, complete, and a lasting contribution to existing knowledge.

AERONAUTICAL NOTES (AN): Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

LABORATORY TECHNICAL REPORTS (LTR): Information receiving limited distribution because of preliminary data, security classification, proprietary, or other reasons.

Details on the availability of these publications may be obtained from:

Publications Section,
National Research Council Canada,
National Aeronautical Establishment,
Bldg. M-16, Room 204,
Montreal Road,
Ottawa, Ontario
K1A 0R6

ÉTABLISSEMENT AÉRONAUTIQUE NATIONAL

PUBLICATIONS SCIENTIFIQUES ET TECHNIQUES

RAPPORTS D'AÉRONAUTIQUE

Rapports d'aéronautique (LR): Informations scientifiques et techniques touchant l'aéronautique jugées importantes, complètes et durables en termes de contribution aux connaissances actuelles.

Rapports de génie mécanique (MS): Informations scientifiques et techniques sur la recherche externe à l'aéronautique jugées importantes, complètes et durables en termes de contribution aux connaissances actuelles.

CAHIERS D'AÉRONAUTIQUE (AN): Informations de moindre portée mais importantes en termes d'accroissement des connaissances.

RAPPORTS TECHNIQUES DE LABORATOIRE (LTR): Informations peu disséminées pour des raisons d'usage secret, de droit de propriété ou autres ou parce qu'elles constituent des données préliminaires.

Les publications ci-dessus peuvent être obtenues à l'adresse suivante:

Section des publications
Conseil national de recherches Canada
Établissement aéronautique national
Im. M-16, pièce 204
Chemin de Montréal
Ottawa (Ontario)
K1A 0R6

UNLIMITED
UNCLASSIFIED

AN OVERVIEW OF TECHNOLOGY FOR SPOKEN INTERACTION WITH MACHINES

UNE INTRODUCTION À LA COMMUNICATION VOCALE AVEC LES MACHINES

by/par

M.J. Hunt

National Aeronautical Establishment

OTTAWA
FEBRUARY 1988

AERONAUTICAL NOTE
NAE-AN-50
NRC NO. 28714

S.R.M. Sinclair, Head/Chef
Flight Research Laboratory/
Laboratoire de recherches en vol

G.F. Marsters
Director/Directeur

Summary

This report provides a non-mathematical introduction to speech input and output technology. It is divided into three parts. The first presents necessary background information on speech: on its nature, its production and perception, and on methods of analysis and coding used in speech I/O. A central message is that our subjective impression of speech is misleading and causes us to underestimate the complexity of speech communication. The second part is concerned with speech output and discusses the trade-offs that must be made between the quality and flexibility of the speech generated and the complexity and storage requirements of the speech output system. The final – and longest – part of the report deals with speech recognition. Arguments are presented in favor of statistical rather than rule-based approaches to speech recognition. The categories of recognizer currently available and the algorithms they use are briefly described, with the general conclusion that the performance obtained depends critically on the training process: on the type and quantity of the training material and on the amount of information derived from it. Three more detailed sections cover spectral representations and distance measures, the particular set of representations classed as auditory models, and techniques for handling noise and distortions. The last section discusses the difficulties of specifying recognizer performance, and recommends that all performance measurements should be treated with circumspection.

Résumé

Le rapport constitue une introduction non-mathématique à la technologie de la parole. Il se divise en trois parties. La première présente de l'information de base sur la parole: sur sa nature, sa production, sa perception, ainsi que sur les méthodes d'analyse et de codage pertinantes à la reconnaissance automatique et à la synthèse de la parole. Une des idées maîtresses de cette partie est que notre impression de la parole ne correspond pas à la réalité, et qu'elle peut nous mener à sousestimer la complexité de la communication parlée. La deuxième partie porte sur la synthèse, et met en evidence les compromis qu'il faut faire entre d'une part la qualité et la flexibilité de la parole générée et d'autre part la complexité et les besoins de mémoire du système utilisé. La dernière – et la plus longue – des parties du rapport est consacrée à la reconnaissance de la parole. Elle se déclare en faveur des approches basées sur les méthodes statistiques plutôt que sur l'application des règles. Une section de cette partie décrit brièvement les classes de systèmes de reconnaissance et les algorithmes qu'ils utilisent, avec la conclusion générale que la performance d'un système est largement determinée par son processus d'apprentissage: par le type et la quantité du matériel d'apprentissage ainsi que par le montant d'information qui en est extrait. Trois sections plus spécialisées traitent des représentations spectrales, des modèles auditifs, et des techniques pour combattre le bruit et les distortions. La dernière section explique pourquoi il est difficile de spécifier la performance d'un système de reconnaissance et conseille qu'il faut traiter toute mesure de performance avec circonspection.

CONTENTS

Section		*Page*
1.	Introduction	1
2.	The Nature of Speech	1
3.	Speech Production	3
4.	Aspects of Speech Perception	4
5.	Speech Analysis and Coding Techniques	6
6.	Techniques for Speech Output	9
7.	Approaches to Automatic Speech Recognition	12
8.	Categories of Recognition Systems	13
9.	Algorithms for Speech Recognition	15
10.	Spectral Representations and Distance Measures	18
11.	Auditory Models	20
12.	Noise and Distortions in Speech Recognition	21
13.	Specifying Recognizer Performance	23
14.	Concluding Remarks	26
	Acknowledgments	27
	References	28

1. Introduction

This report is intended to provide non-specialist readers with a framework in which to view speech recognition and synthesis technology. It makes no claim to be a complete, scholarly account, and the references are simply pointers into the technical literature: citing a particular publication does not necessarily imply the precedence or special importance of that work. Algorithms are not described in detail and mathematics is avoided. Some parts do, however, assume a general technical background. The presentation tries to keep a balance between obscurity and the misleading oversimplification that is common in popular accounts of the field. Readers must decide into which trap it has fallen.

The report can be divided into three parts. The first, spanning sections 2 to 5, provides background information on speech necessary for understanding the rest of the report. The second part, in section 6, discusses options for speech output. The third – and longest part – covers speech recognition. Applications-related issues (such as error-correction strategies) and approaches to language modeling are not discussed in detail.

Coherent surveys inevitably take a particular point of view; but wherever an issue is clearly contentious, I will try to point out that I am expressing my personal opinion.

2. The Nature of Speech

Our internal impression of speech is misleading. Our perception of its being composed of discrete, immutable words, themselves composed of discrete, immutable speech sounds – often mistakenly called *phonemes* – corresponds in no way to the properties of the acoustic signal. What we perceive is the output of a sophisticated message reconstruction process [1], much as when we see a picture we immediately reconstruct a three-dimensional scene from it. In reconstructing the scene we unconsciously use our knowledge of the world, of lighting effects and of perspective. In the same way, in reconstructing a spoken message we unconsciously use our knowledge of the language – of its grammar and phonetics – of the situation, of the speaker, and of the world in general. This reconstruction is so automatic and effective that we are mostly unaware that we are continually having to supplement the information present in the acoustic signal: we only realise that the "s," "f" and "th" sounds in *lass, laugh* and *lath* are indistinguishable on the telephone when we have to spell out an unfamiliar name. The inadequacy of the information in the acoustic signal and the need for a wide knowledge of the world in "making sense" of the acoustic information mean that automatic speech recognizers that are as skilled as humans at recognizing unrestricted speech are a distant prospect.

There are no gaps between spoken words; indeed, there are no consistent cues to word boundaries. Our impression to the contrary is a result of the message reconstruction process coupled with our familiarity with written language, which itself reflects an internal representation. Words interact at their boundaries: ***bread board*** is often pronounced as "breab board," and the vowel in ***two*** in the sequence ***two six*** is often close to the vowel in French *tu*. Short, low-content words like ***a*** or ***of*** can be unrecognizable when excised from their context.

The idea that speech sounds resemble written letters is even more misleading. Some consonant sounds – "d," "b" and "g," for example – cannot be pronounced in isolation: they must be preceded or followed by another sound, usually a vowel. Moreover, we cannot take a recording containing one of these sounds and cut away the surrounding speech until the sound is heard recognizably in isolation. The cues to the identity of such sounds – and to many others – are in the continuous transitions into the adjacent sounds. They do not exist as discrete acoustic entities. Even vowels, which certainly can be produced and perceived in isolation, can sound quite different when excised from the middle of a word. More often than not, there is no discernible boundary between one speech sound and its neighbors.

The impression that words are made up of sequences of discrete, immutable sounds seems to come from a high level in our processing of speech. At this level words appear to be coded as sequences of units called *phonemes* [2] that are capable of specifying all the distinctions between words that can be made in the particular language. Whether the phoneme has an objective existence in the brain has yet to be established, though it seems likely, and in any case it is a natural and useful abstraction in describing spoken language.

In conjunction with a set of production rules, phonemes provide a specification of how a given word is to be pronounced. The production rules are context sensitive: for example, the /t/ phoneme in English results in phonetically and acoustically different sounds when it begins or ends a word, as in *tap* and *pat,* when it is preceded by an /s/, as in *stick* or followed by an /r/, as in *tree.* We are generally unaware of these differences because they cannot change the meaning of a word. If some of these differences did correspond to phonemic distinctions – as they do in many other languages – then they would be much more noticeable. Nasalization in vowels, for example, is a phonemic cue in French and is therefore highly noticeable, while in English – where it is at least as common – it is not a phonemic cue and is therefore largely unnoticed.

Some of the phonetic cues used to signal phonemic distinctions can be surprising. For example, the words *ones* and *once* differ in their final phoneme, /z/ and /s/ respectively. But in normal speech the main phonetic distinction between them is in the length on the "n" sound, which is typically twice as long in *ones.*

All the information in a written message is contained in its words, and the words themselves are completely specified by the letters from which they are composed. A spoken message, on the other hand, contains more information than simply the sequence of words and the sequences of phonemes from which the words are composed [3]. Apart from side information concerning the identity and general emotional state of the speaker, there is much additional information concerned directly with the meaning and structure of the message. This information is encoded in *prosodic* features: pitch, loudness and timing. Through the intonation and rhythm of a sentence we can deduce its grammatical structure, as well as information such as whether the speaker intends to continue speaking. We can often infer something of the speaker's attitude: a sentence as short as the word *yes* can convey scores of different meanings depending on the way it is said.

New or important information is also highlighted by prosodic cues: "the *new* red car" and "the new *red* car" mean different things.

Words in English also have inherent prosodic properties in the level of stress that is assigned to each syllable. Thus, the noun *permit* and the verb *to permit* are pronounced differently even though they are composed of the same phonemes. Since prosodic cues signal information at many different levels — word identity, sentence structure, speaker attitude, *etc.* — they are difficult to analyze, and no practical speech recognition system makes use of them. Techniques have instead been developed to allow prosodic information to be ignored in identifying sequences of words. For most current uses of recognizers — simple commands, recording strings of digits, and so on — this strategy is adequate, though if we are ever to approach human levels of performance, it will not be.

Human listeners, on the other hand, cannot ignore prosodic information. It is therefore essential that speech output devices generate speech with appropriate prosody. When we are presented with a synthesized sentence in which the words are pronounced as though they were spoken in isolation, each word is perfectly clear, but the sense of the sentence is almost impossible to retain.

3. Speech Production

To a good approximation, the acoustic process of speech production can be modeled as a source driving a linear filter with little interaction between the source and the filter. The amplitude spectrum of a speech sound is therefore the product of the source spectrum and the amplitude response of the filter.

In the most important class of speech sounds, namely *voiced* sounds, which includes the vowels, and many consonants such as /l/, /m/, /d/, the source is provided by the vocal cords. The vocal cords open and close at a rate of around 100 Hz in men and closer to 200 Hz in women. This rate, which changes only slowly, is known as the *fundamental frequency* or F_0, and it correlates strongly with the perceived pitch of a speech sound. When we sing a tune, it is the fundamental frequency that follows the notes.

Most of the acoustic excitation in voiced sounds is concentrated at the instant of closure of the vocal cords, and a good approximation to voiced excitation after correcting for an overall 12 dB/octave roll-off is provided by an impulse at the instants of closure. With a steady fundamental frequency, the spectrum of the excitation is therefore a set of harmonics separated from each other by the fundamental frequency and decreasing in amplitude by 12 dB/octave. Deviations from the impulse idealization [4] lead to changes in the quality of the voice — giving it a tense, breathy or falsetto quality, for example — but in western languages such changes do not alter the phonetic content of the sound.

In *voiceless* sounds (consonants such as /p/, /f/ or the /s/ in *sea*) the excitation source is provided by a constriction in the vocal tract where turbulent airflow is generated — in /f/, for example, the constriction is between the lower lip and the upper teeth. This excitation has no periodic component, and can be modeled as white noise.

The third and least important class of sounds, including /v/ and /z/, have *mixed excitation.* That is, the source consists of noise excitation from a constriction as well as periodic excitation from the vocal cords.

The soft palate at the back of the mouth acts as a valve determining whether air can flow through the nasal cavity and out through the nostrils or not. In *nasal consonants* (/m/, /n/ and the "ng" sound in *sing*) the valve is open but airflow through the mouth is shut off. In *nasalized* sounds, on the other hand, air flows through both the oral and nasal cavities. We have already noted that although nasalization is not a cue to phoneme identity in English it is nevertheless common. The vowel in a word such as *man* is almost always nasalized.

In non-nasalized sounds the vocal tract is configured as an unbranched tube. At the frequencies important for speech, sound propagation in this tube is effectively planar, and reflections in the tube are determined by changes in its cross sectional area. The filtering effect of the tube can be represented as a sequence of resonances, *i.e.* as an all-pole filter, and the frequencies and bandwidths of these resonances -- known as *formants* -- are determined by the positions of the tongue, lips and jaw. These parameters, together with the roughly 6 dB/octave high-frequency lift caused by radiation from the mouth, completely specify the filtering effect of the vocal tract in oral sounds.

The lowest frequency resonance, known as the first formant or F_1, has an average value of around 500 Hz in men and varies in the range from 300 to 700 Hz. Succeeding formants are spaced on average 1 kHz apart. In women, formant frequencies are on average 15% higher than in men, since their vocal tracts are typically 15% shorter. The higher formants generally have larger bandwidths.

In voiceless sounds, there is often too little low frequency energy to excite F_1 appreciably. The remaining formants have broader bandwidths than in voiced sounds, resulting in smooth, featureless power spectra.

In nasal and nasalized sounds, the vocal tract becomes a branched tube. The branching gives rise to antiresonances and thus to a transfer function containing zeroes as well as poles. Such sounds also show additional resonances, and bandwidths are increased.

4. Aspects of Speech Perception

Human auditory perception [5,6] resembles that of other animals and has therefore not adapted significantly to deal with speech. It is likely, on the other hand, that speech has evolved to suit the properties of human auditory perception. Features of the acoustic signal that are imperceptible obviously cannot be useful to human listeners. Any information they carry is accidental and is unlikely to be well controlled. Studying auditory perception should therefore give us clues to the important cues in speech, and modeling perception will probably lead to effective representations for automatic speech recognition.

The subjective loudness of a sound is not directly proportional to its acoustic power. Energy at low frequencies (below about 200 Hz) and at high frequencies (above about 4 kHz) counts for less than energy at intermediate frequencies.

Also, equal increments in loudness correspond more closely to equal power increments measured on a log scale than on a linear scale.

The ear is generally considered to be insensitive to the phase structure of a sound. In fact, if two spectral components are close enough in frequency (within about 100 Hz at low frequencies, more at high frequencies) their phase relationship is noticeable. Nevertheless, phase insensitivity over larger frequency differences means that pairs of waveforms can look quite different and yet sound indistinguishable. This is why analysis methods usually represent the short-term power spectrum of the speech signal and ignore its phase spectrum. Phase insensitivity is important for human speech perception because room reverberation and changes in voice quality affect the phase spectrum and would otherwise cause speech to sound different.

The frequency resolution of the ear is not uniform across the spectrum; rather, it decreases at higher frequencies. It is often approximated by a scale that is linear up to 1 kHz and logarithmic from then on, with the range 0-1 kHz being considered perceptually equivalent in size to the range 1-4 kHz. Perceptual frequencies are measured on the *mel* [7] or *bark* [8] scales.

As one would expect from signal theory, the lower frequency resolution at high frequencies is coupled with better time resolution. This property suits speech well, since the sounds that need to be distinguished by fine time resolution, such as the plosive/fricative pairs /t/ and /s/, have their energy concentrated at high frequencies and have little spectral fine structure, while sounds such as the vowels that need to be distinguished by details of their power spectrum have most of their power concentrated at low frequencies.

Loud sounds suppress the ear's response to succeeding sounds at the same frequency. Thus, a loud tone can mask the presence of a similar weaker tone presented just after it. This phenomenon is known as *temporal masking*. The masking effect decays with time over a period of 100 ms or so. It enhances the perceptual salience of onsets and of spectral changes, such as formant transitions between consonants and vowels.

A second kind of masking occurs between components at different frequencies presented at the same time. It is therefore known as *simultaneous masking*. The response to a weak tone can be suppressed by the presence of a nearby strong tone, hence this phenomenon is also called *two-tone suppression*. The amount of suppression decreases as the tones move farther apart, and the decrease is much faster when the masking tone is higher in frequency than the tone it is masking. Thus, simultaneous masking operates mainly upwards in frequency.

Simultaneous masking is probably responsible for several properties observed in the perception of speech sounds. The most obvious of these properties is our ability to ignore low levels of background noise: with wideband noise, intelligibility is largely unaffected until the signal-to-noise ratio falls below 15 dB.

Many of the masking-related properties concern formants. Experiments show that we are extremely sensitive to the frequencies of formants but relatively insensitive to their bandwidths. When pairs of formants come close to each other

(within about 300 Hz at low frequencies) they fuse and cannot be distinguished from a single equivalent formant. This fusion occurs over much larger frequency differences than would be expected from our ability to detect changes in the frequency of a tone, but they are consistent with the range over which two-tone suppression operates.

Simultaneous masking may also explain why we are largely insensitive to the details of the fourth and higher formants in voiced sounds. However, the relatively low variability of the higher formants on a perceptual frequency scale may also help to explain their weak influence on phonetic distinctions.

When listeners are asked to judge the *phonetic* similarity of two speech sounds they seem to use different criteria from those they use to judge the overall similarity of the stimuli simply as sounds [9]. For phonetic judgments listeners seem able to ignore large differences in spectrum balance (the smooth spectral shaping that, for example, the tone controls on an audio amplifier affect) [10]. An ability to ignore spectrum balance differences is useful, because differences in speaking level affect the spectrum balance as do the reverberant properties of rooms. Moreover, without this ability communication over the telephone would be all but impossible.

Finally, there are two properties involving fundamental frequency that are worth mentioning. The first is our ability to perceive – to *hear* – the fundamental frequency of a speech sound even when the fundamental itself has been filtered out leaving only its higher harmonics. This property is crucial for the intelligibility of speech over long-distance telephone lines, where the bottom 300 Hz may be missing.

The second such property concerns our perception of the first formant, especially when the fundamental frequency is high. The frequency resolution of the ear at low frequencies is fine enough to allow individual harmonics of the fundamental to be detected. Nevertheless, in making phonetic comparisons between pairs of speech sounds with different fundamental frequencies listeners do not seem to match the strongest harmonic in the F_1 region; rather, they seem able to deduce the frequencies of the formant in the two sounds and base their judgment on that. There is some question whether listeners ignore the frequencies of the harmonics completely [11], but it seems clear that to a substantial extent they do.

In summary, judgments of the phonetic identity of voiced speech sounds depend heavily on the frequencies of the first two or three formants. These judgments are affected relatively little by the higher formants, by details of formant bandwidths or amplitudes, by phase properties, or by fundamental frequency.

5. Speech Analysis and Coding Techniques

The properties of speech production and perception described in the last two sections largely determine the approaches to speech analysis and coding.

Since the parameters of the acoustic source and filter in speech production vary slowly, it is efficient to separate the source and filter and code their parameters separately. Such parameters need only be estimated fifty or a hundred times

a second, while if the resulting waveform is encoded, it must be sampled at least eight thousand times a second. Because of phase insensitivity, only the amplitude response of the filter need be estimated, while the source can be described by its amplitude, a voicing decision, and the fundamental frequency in voiced speech. Speech coding systems that separate source from filter are known as *vocoders*. For speech recognition purposes, as opposed to speech coding and synthesis, only the filter characteristics – *i.e.* the smoothed short-term power spectrum or *spectrum envelope* – are of interest.

The first vocoders were *channel vocoders* [12], in which the spectrum envelope is estimated by measuring the energy in a bank of band-pass filters. The channels are spaced farther apart at higher frequencies to reflect the frequency resolution of the ear. Channel vocoders can be entirely analogue devices, though today they are usually implemented using digital filters or FFT's.

The second major class of vocoders are *linear predictive vocoders*, or *LPC* (for "linear predictive coding") devices [13]. LPC assumes that the vocal tract can be modeled as an all-pole filter and that after preemphasis the source in voiced sounds can be modeled as a sequence of impulses. In so far as these assumptions hold, and given that the number of resonances in the filter is known, LPC can in principle determine the filter parameters exactly by analyzing the autocorrelation properties of the speech waveform within a single excitation cycle.

We have seen that the all-pole assumption is not valid for all speech sounds. However, even when it is invalid – in nasal sounds, for example – LPC can still do a reasonable job of estimating the spectral envelope. In these cases, it simply fits to the spectrum the closest envelope it can find that would be generated by an all-pole filter of the given order. Conveniently, the spectrum fitting is not based on the usual least-squares criterion, but rather on a criterion that concentrates on fitting the high-energy regions of the spectrum at the expense of the weak regions. This criterion follows naturally from the fact that the LPC analysis does a least-squares fit to the waveform, since the high-energy parts of the spectrum affect the waveform most. It reflects to some extent the simultaneous masking properties of the ear, which also lead to a concentration on the high-energy parts of the spectrum. On the other hand, LPC does not easily lend itself to reflecting the varying frequency resolution of the ear in the way that channel vocoders do.

A more serious weakness of LPC as it is usually implemented comes from its breaking the requirement that the autocorrelation analysis should be carried out over an unexcited portion of the waveform. Since it is difficult to determine reliably from the waveform just when the excitations occur, the analysis is instead carried out over a fixed-length window covering several excitation cycles. Moreover, since the use of the exact *covariance* method of LPC would in these circumstances often result in the LPC analysis specifying unstable filters, the simpler but inexact *autocorrelation* method is used. This latter method guarantees filter stability at the expense of precision in estimating the parameters of the resonances.

For speech coding purposes some of the weaknesses of LPC can be alleviated by encoding additional information specifying a more complex time varying excitation function. *Residual excited linear prediction* (RELP) [14] and the more recent and popular *multipulse* LPC [15] are examples of this class of coding systems. Such variants have no effect on the quality of the filter specification, and have consequently little to contribute to speech recognition.

Most of the compromises made by designers of LPC systems are imposed by the requirement for real-time operation in communications systems. For many applications of speech analysis – in particular for speech output systems – there is no need for the analysis to be done in real time. In these circumstances, *pitch-synchronous LPC* [16] can be considered. This approach has been studied intensively in our laboratory [17]. Following the theory of LPC, the autocorrelation properties are computed over a window placed between consecutive excitation instants in voiced speech. The excitation points are determined using a device called a *laryngograph* or *electroglottograph* [18], which measures the radio frequency impedance across the larynx and hence the area of contact of the vocal cords. If the laryngograph signal is recorded in parallel with the speech signal, it can be used to determine the instants of closure of the vocal cords and thus the instants of excitation of the vocal tract. The use of the laryngograph also provides a far more reliable measure of voicing and fundamental frequency than is possible using only the speech signal. Errors in voicing and fundamental frequency are a major source of degradation in speech resynthesized from an LPC analysis. Also, the exact covariance method of analysis can be used, with the resulting infrequent instabilities being detected and replaced by equivalent stable configurations.

Pitch-synchronous LPC is more sensitive than conventional LPC to the quality of the speech signal. The recordings must be made in extremely quiet, non-reverberant conditions and digitized with care. For non-real-time applications, however, the extra effort is well justified by the accuracy the formant analysis obtained and the high quality of the resulting resynthesized speech. The reliable formant analysis allows the voice to be manipulated in ways that are impossible with conventional LPC. Moreover, unlike other techniques for improving the quality of LPC speech, such as multipulse and RELP, pitch-synchronous LPC does not increase the amount of data needed to resynthesize the speech, and resynthesis can be carried out using standard LPC synthesis hardware.

Many techniques have been developed for coding the speech waveform without separation of source and filter information. These techniques result in higher data requirements than the vocoder approaches. As they are peripheral to the rest of the discussion here, they will not be discussed further. Equally, discussion of speech analysis using auditory models and so-called perceptually based LPC is deferred until the sections on speech recognition.

Instead of encoding individual samples of a speech waveform, a sequence of such samples can be encoded by classifying it as a member of a particular group represented by a reference sequence out of a codebook. The codebook usually contains several hundred such groups. The information on the sequence to be transmitted or stored is then simply the index of the reference sequence selected.

This process is called *vector quantization* [19]. To work well, the classes represented by the reference sequences should divide up the space in a way that reflects both the perceptual discriminability and the probability of occurrence of waveform shapes. Thus, common sequences should be encoded more accurately than rare ones by having a higher density of classes in the common regions, while the perceptual differences between adjacent reference forms should be reasonably uniform over the space. The technique has been extended to encoding the sets of channel energies across the filter bank in a channel vocoder and the sets of filter coefficients specifying one analysis in an LPC vocoder.

6. Techniques for Speech Output

The possibilities for speech output range from *text-to-speech* systems [20] capable of taking any piece of text in the language and generating a spoken version of it to systems that are little more than recording devices echoing back the fixed spoken phrases that are stored in them. The parameters controlling the choice of system appropriate for a particular application include the intelligibility and naturalness of the speech, the complexity of the system, and the range of vocabulary, sentence structures and voice types that need to be generated. Even when the vocabulary size is not enormous, text-to-speech systems offer the considerable advantage of allowing the addition of further words to an existing system. Adding extra words to a system that uses a particular person's voice, on the other hand, entails either having continued access to the person or re-recording the whole vocabulary.

The weak connection between spelling and pronunciation presents an obvious difficulty for text-to-speech systems in English. However, by use of pronunciation rules together with long lists of exceptions, and in some cases techniques for dividing up words into sub-units called *morphs* (*un-*, *re-*, *-able*, *-ing* and *-ly* are morphs), the systems are able to find a correct phonetic transcription for most words.

Working out the prosodic features of a sentence is much harder. The stress levels to be assigned to syllables in a word can be found in the same way as its pronunciation, but as we saw in the first section the timing and intonation and loudness contours of a sentence depend on its grammatical structure, its meaning and on the attitude and emotional state of the speaker. Sophisticated systems can usually analyze the grammatical structure of a sentence, but the other factors depend on an understanding of the text and ultimately of the world. Perfect speech synthesis from unrestricted text is about as far away as perfectly accurate speech recognition of unrestricted material.

Our cultural emphasis on the written form of language has led us to ignore those aspects of speech that are not directly reflected in text. This has resulted in speech output being seen as nothing more than a plug-in replacement for a text display, with a consequent underexploitation of its potential [3]. Every time someone speaks, much information is transmitted about the sex, geographical origin, identity, and the emotional and sometimes even the physical state of the speaker. Yet almost none of this information is used in speech output systems. One of the best current text-to-speech systems, *Dectalk*, does, it is true, allow for

some fairly crude voice changes, but I do not believe that the feature is widely used. Friendly reminders should sound friendly and – at least in some circumstances – urgent warnings should sound urgent, while different sources of information should be linked to recognizably different voices.

Waveform-Based Speech Output

The simplest speech output systems replay the digitized speech waveform. As CD players show, these systems can generate speech that is indistinguishable from the original. The cost in storage requirements, however, can be extremely high – anything from 64 kbits/s to 700 kbits/s, depending on the quality needed. By using more sophisticated waveform coding techniques the storage requirements can be cut to as low as 12 kbits/s, but there is a gradual trade-off between bit rate and quality, and the complexity of the hardware is increased.

The biggest disadvantage of waveform coding is that it allows virtually no flexibility in the output: words and phrases have to be replayed exactly as they were recorded. Even with messages as simple as lists of digits, there are substantial differences in timing and intonation between final and non-final items. Words must therefore be recorded in all the contexts in which they are to be used. Applications with a small number of fixed messages are well suited to this requirement. To generate spoken telephone numbers in response to requests for directory assistance, some telephone companies concatenate waveform-encoded digits recorded in a range of contexts.

Synthesis-by-Rule

At the other pole from waveform coding lies synthesis-by-rule [21]. Words to be synthesized are first transcribed as sets of phonetic symbols and rules are then applied to compute trajectories of formant parameters. This is certainly the most complex approach to speech output. On the other hand, for large vocabularies it has by far the lowest storage requirements. It also possesses the greatest flexibility, allowing both the prosodic and phonetic features of words to be altered according to the context and offering the possibility of multiple voices.

The worst synthesis-by-rule systems can be unintelligible unless the content of the message is known beforehand. Manufacturers of such systems often use well-known passages such as the Gettysburg address to hide the low intelligibility. We saw in the first section how inadequate acoustic information is unconsciously supplemented by the listener's knowledge. Casual, subjective impressions of intelligibility are unreliable.

Even the best synthesis-by-rule systems generate voices that sound distinctly inhuman. It can be argued that speech output from machines should not sound human, since it invites listeners to confer expectations of human intelligence on the machine [22]. However, the inhuman nature of synthesis-by-rule speech probably reduces its intelligibility. Although such speech scores moderately well when the intelligibility of isolated words is measured, human response times to such speech are found to be significantly slower than to natural speech, and comprehension of passages is lower [23].

Because of their high cost, low naturalness, and questionable intelligibility, synthesis-by-rule systems have so far found few applications. They are appropriate in applications such as text readers for the blind, where a large vocabulary is essential and the highly motivated users will tolerate the unnatural quality. They are also appropriate for regular professional users who can become familiar with the voice and a restricted range of messages and eventually cease to notice the poor quality of the speech. In the long-term, synthesis-by-rule will surely be improved to the point where it is the system of choice for almost all applications.

Vocoder-based Speech Output

Intermediate between the extremes of waveform coding systems and synthesis-by-rule lie systems using approaches originally developed for low-bit-rate digital speech communications systems. Vocoders typically operate at around 2.4 kbits/s, so the storage requirements as well as the complexity and the naturalness and intelligibility of the speech is intermediate between the two other approaches just described. Prosodic features can be manipulated, allowing the same word to be used in multiple contexts and thus further reducing the effective storage requirements. Usually, however, phonetic features cannot be easily modified, so word-boundary interactions cannot be introduced, and voice characteristics cannot be manipulated. The Texas Instruments toy *Speak 'n Spell* was a pioneering example of the use of this kind of speech output in consumer products.

As mentioned in the previous section, the quality of LPC vocoder speech can be improved at the expense of bit rate and complexity by using techniques such as RELP and multipulse LPC. Systems of this kind operate at around 9.6 kbits/s. However, for speech output applications they suffer from a serious drawback in that the fundamental frequency contour cannot be easily changed. They are therefore as rigid as waveform coding systems. Compared with waveform coding, they offer reduced storage requirements at the price of increased complexity.

Pitch-Synchronous LPC

As we saw in the previous section, pitch-synchronous LPC offers a considerable improvement in the quality of LPC speech without any increase in storage requirements or in the complexity of the synthesis system. Moreover, in contrast to other methods of improving LPC quality, pitch synchrony actually increases flexibility. The accurate formant analysis means that not only prosodic features but also phonetic and voice quality features can be manipulated. Apparent sex and dialect changes are possible, as well as the introduction of emotional attributes such as tense and trembling voices and coarticulation phenomena at word boundaries.

Diphone and Demisyllable-based Synthesis

Diphones and demisyllables offer an alternative to synthesis-by-rule in constructing text-to-speech systems. Diphones attempt to capture the important transition information between consecutive sounds in units that pass from the steady portion of one sound to the steady portion of the next [24]. Demisyllable

units [25] are somewhat longer, allowing for the consonant clusters that can occur in the first and second halves of syllables. In both cases, the units are generally encoded using LPC, and by concatenating sequences drawn from an inventory of around 1000 such units, unrestricted text can be generated. The method has the flexibility of the LPC system that was used to produce the units, and the units contain the degradations introduced by the LPC system. The use of pitch-synchronous LPC therefore seems logical. There is a trade-off between quality and storage requirements, since a larger number of units allows context and vowel reduction phenomena to be modeled better. Although diphone synthesis seems attractive, I am not aware of any practical applications in English of such systems.

7. Approaches to Automatic Speech Recognition

Someone once said that speech synthesis and speech recognition were complementary problems: synthesis was like squeezing toothpaste out of the tube, and recognition was like trying to put it back in again. The analogy understates the difficulty of creating good synthetic speech, but there is some truth in it. In synthesis, it is sufficient to generate a single acceptable form of a phrase; in recognition, on the other hand, it is necessary to cope with a range of different but nevertheless normal versions of any word or phrase. Speech recognition therefore amounts to looking for invariant features linking different examples of the same speech and distinguishing them from examples of different speech.

Approaches to speech recognition can be largely divided into two broad classes. The first attempts to set up a system of rules embodying human knowledge of what characterizes speech sounds as they appear in spectrograms: an *expert system* modeling a skilled spectrogram reader. The second, by contrast, uses little human knowledge, but rather attempts to use statistical properties of training material in systems comparing patterns on continuous scales rather than applying tests with binary outcomes.

Proponents of the first, knowledge-based, approach sometimes claim it to be more intelligent and sophisticated than the statistical approach. In my judgment, the statistical approach has, nevertheless, been more successful up to now; and most recognition products – indeed *all* recognition products that have been extensively tested – fall in this camp. A few years ago, the popularity and success of expert systems in other domains lead to intense interest in their application to speech recognition, but this interest has since diminished considerably. Currently, there is a major U.S. DoD DARPA program in speech recognition with two large, competing teams, one at BB&N [26] taking a statistical approach and the other at CMU [27] taking a knowledge-based approach. Initial results are showing a strong advantage for the statistical approach, and – perhaps more remarkably – a single graduate student at CMU is reported to have developed a statistically based system that is massively outperforming the knowledge-based system developed by the large professional team [28].

We have seen that our internal impression of speech is misleading, and the way in which we understand speech is not open to conscious examination. Attempts to understand speech from spectrograms constitute an attempt to

decode the speech signal by conscious reasoning. Consciously designed signals – printed text, Morse code, teleprinter transmissions *etc.* – typically have easily identifiable basic units into which the signal can be unambiguously segmented. Speech does not have such units and does not seem to be the kind of signal that the conscious mind would design, nor that it would be good at decoding [29]. Consequently, it is not surprising that the most practiced and adept spectrogram readers [30] are much less effective at decoding speech than the least adept listeners with normal hearing. *Expert systems* are intended to model human conscious reasoning, and attempts to use such an approach to decode speech information at the phonetic level by modeling the behavior of a phonetician reading spectrograms seem destined to be less successful than attempts to model *unconscious* speech perception processes. Statistical methods do not explicitly attempt to model speech perception processes, but since speech has probably been shaped to meet the needs of human perception, its statistical properties are likely to reflect these needs.

This is not to say that expert systems do not have potential for speech recognition: the approach may be appropriate for the organization of higher level syntactic and particularly semantic information, which *is* susceptible to conscious analysis. The effective use of such higher level information will be necessary if we are to achieve really sophisticated speech recognition. However, since the interest here is in what can be done now, the emphasis will be on statistical methods.

8. Categories of Recognition Systems

The simplest recognizers – *isolated-word* devices – accept words or fixed phrases surrounded by periods of silence. Generally, an energy criterion is used to decide where a word starts and stops. The signal between these instants is then matched against stored word patterns. Except for a few hybrid systems, isolated-word recognizers require the periods of silence to be longer than any silent periods that may occur before stop releases within words, namely about 200 ms. Thus, apart from forcing the user to adopt an unnatural style of speech, isolated-word systems slow down the rate at which speech can be input.

Systems that accept continuous input are called either *continuous* speech recognizers or *connected-word* recognizers. Many writers make a distinction between the two terms, but the distinctions are not consistent. Most such systems do not take into account the interactions between words. Thus they effectively recognize sequences of isolated words spoken in a connected manner. The term *connected-word* is used to signal this fact. Alternatively, the term may be used to distinguish early systems that required the user to pause after a certain length of input from more recent systems that allow the user to speak without ever pausing and output the sequence of words recognized with some variable delay, typically a couple of words. The term *connected-word* will be used here in the first sense.

Even when the input to a recognizer is to take the form of isolated words, a connected-word recognizer may be the better choice for the task. Connected-word recognizers explain all the input as a sequence of words and various non-word reference patterns representing silence and perhaps background noises, breath

noises, lip smacking *etc* [31]. This makes them better able to cope with extraneous noises than an isolated-word system, whose energy-thresholds may be inappropriately triggered.

In addition to isolated and connected-word modes, recognizers may work in *word-spotting* mode. In this mode, the stored word patterns may be matched between any two points in the input without the requirement that they should fit into some explanation for the complete input. Originally developed for intelligence applications, this mode is often incorporated in connected-word recognizers, where it can be used to allow the recognizer to scan its input for a key-word that turns on its normal recognition mode. Such a feature allows a user to break off to talk to other people without risking spurious responses from the recognizer or having to push a button to inactivate it.

There are major differences between recognizers in how they are trained. Statistically based recognizers are trained by being given examples of the words, or, more rarely, sub-word units, that make up the vocabulary. The simplest recognizers accept a single example of each word from the speaker who is to use the device. More sophisticated recognizers require multiple examples, which they align and average together; while the most sophisticated devices compute not only the average spectral and timing properties, but also their variances. The ability to accept large amounts of training material – and to make good use of it – is probably the biggest single factor in determining recognizer performance.

Connected-word recognition poses a special problem because words tend to be pronounced differently when spoken continuously and in isolation. Most connected-word recognizers simply use an isolated-word training procedure, using an energy-threshold to determine word end-points. Some devices, however, having made initial reference forms from isolated words then go on to a second stage in which these isolated-word examples are used to pick out examples of the same word in continuously spoken word sequences. This *embedded training* seems to be useful in deriving more representative reference forms for continuously spoken words.

Speaker-independent systems are systems that are trained by speakers other than the current user. Generally, many speakers would be used for the training and there are often several reference forms for each word [32]. Speaker independence generally results in a reduction in recognizer performance, and there will always be some speakers whose speech is too far from the population norm to allow them to use a particular system. It is a particularly desirable quality for applications where individuals use the system only briefly and where the vocabulary is large.

Manufacturers have sometimes claimed speaker dependence as a positive feature because it would prevent unauthorized use. This seems like claiming that small, cramped cars have the advantage of being unlikely to be stolen by tall thieves. But it is also based on the misconception that there is a clear distinction between speaker-dependent and speaker-independent systems. Any speaker-dependent system can be used by other speakers: there will simply be more recognition errors. Since many of the variations between speakers also occur to a lesser extent within the speech of an individual speaker, good speaker-dependent

systems are likely to be more tolerant of other speakers than bad systems are.

Speaker-adaptive systems [33,34,35] are intermediate between the extremes of speaker dependence and independence in that there is partial retraining to the current speaker. Such systems are particularly attractive in applications where a given speaker will not use the system for long enough to justify complete training, but long enough to allow useful adaptation (which, according to our experience, may be as short as three words). It seems likely that humans use a form of speaker adaptation, at least when presented with an unfamiliar dialect, and in the long-term the most effective recognizers will probably do something similar.

9. Algorithms for Speech Recognition

Statistically based algorithms generally represent the speech to be recognized as a sequence of *frames* containing information about the short-term spectrum. There is typically one frame every 10 or 20 ms. The input frame sequences are compared against reference models.

Models usually represent whole words, but there are advantages in both smaller and larger units than the word. Since speech sounds interact in ways that we are at present poor at characterizing, it is desirable that interactions should be largely contained within the models rather than occurring between them. The larger the unit the more interaction is contained within it. On the other hand, smaller units allow more words to be recognized for a given amount of storage and computational effort. The smallest unit widely used in experimental systems is the demisyllable [36], containing the important transition information between consecutive sounds. The syllable offers an intermediate possibility [37], accounting for a substantial proportion of the speech-sound interactions within words. Beyond the word, units such as word pairs can be considered. Such large units might find a use in improving recognition performance on small vocabularies such as the digits. Whole words, however, are not only a good compromise, but are simplest for training, since it is natural for a user to provide examples of single words.

Most model-based recognizers take a Markov modeling approach to recognition. In *dynamic programming time warping* algorithms [38] the models consist of sequences of frames, called *templates,* and the input sequences are matched against the set of templates. The matching process allows frames in the model or in the input or in both to be repeated or to be jumped over. Repeating or jumping over frames corresponds to stretching or compressing time, which is the origin of the term *time warping.* Given some method, to be discussed in the next section, of determining the similarity of pairs of frames, the dynamic programming algorithm can find the alignment of frames between the input and a given template that has the best total similarity. The template with the best such total similarity is taken to be the most probable interpretation of the input. In connected-word recognizers the criterion is extended to the best sequence of templates that are matched against the input.

The ability to stretch and compress time allows the algorithm to accept inputs with non-linear timing variations. This property is essential because the durations of different parts of words vary widely even when the words are spoken

carefully in isolation, and the variations are much greater in continuous speech, where durations are affected by the prosody of the phrase or sentence. On the other hand, durations of speech sounds play an important role in discriminating between words: we have already seen that *ones* and *once* are mainly distinguished by the duration of the /n/, and, more obviously, the length of the initial noise burst is largely responsible for distinguishing between *tea* and *sea*. To discourage false matches between word pairs distinguished mainly by durations, time distortion is often penalized or limited to a certain maximum amount, say to a 2:1 stretch or compression. Ideally, time warping penalties should reflect the timing variability of the different parts of words, which is presumably lowest in regions where duration carries most useful discriminating information. In practice, dynamic programming systems almost always impose uniform time warping penalties. Equally, spectral similarity measures are generally treated as uniform across words, even though some parts of words are more variable in their spectra than others.

A more recent alternative to dynamic programming time warping is ***hidden-Markov modeling*** [39]. This approach assumes that words can be modeled as a Markov sequence of a small number of hidden states. Each hidden state is probabilistically related to a set of observable states corresponding to the spectral representation of the speech. In producing an example of a word, a hidden-Markov model will stay in the first state for a certain time, generating a distribution of surface states. It then moves to the second state and generates a different distribution of surface states, and so on until the end of the word model is reached. The discovery that made hidden-Markov modeling possible was an algorithm that allows the model parameters to be estimated from a set of training examples of the word. The parameters are the probabilities of staying in a given hidden state or making the transition to the next state, and the probabilities of generating a particular surface state given a certain hidden state. From these probabilities, the probability can be computed that a word to be recognized was generated by the hidden-Markov model for a word in the reference vocabulary.

When hidden-Markov modeling was first introduced, it was believed to be radically different from dynamic programming. However, the two types of algorithm are now increasingly seen to be aspects of the same basic approach [40]. Early formulations of hidden-Markov modeling looked radically different from dynamic programming because of two factors. First, the surface states were discrete, *i.e.* vector quantized. It was shown later that the surface states could be described by continuous distributions, with some increase in computation cost but with an improvement in performance [41]. Second, the log probability that a given model could have generated a word example was originally computed by summing over all possible hidden-state interpretations. This *Baum-Welch* decoding method is theoretically correct and probably gives best performance. It is still the method used at IBM. However, elsewhere it has been widely replaced by the computationally simpler *Viterbi* decoding method, in which the log probability is summed over only the most likely interpretation. Viterbi decoding corresponds to the summation in dynamic programming time warping over the most likely sequence of frame matches.

A central difference between dynamic programming time warping and hidden-Markov modeling is that the former assumes a smooth evolution of the spectrum while the latter makes the less reasonable assumption of abrupt transitions from states with statistically stationary spectra. However, as the number of states in a hidden-Markov is increased from the usual five or so to value closer to forty typical of the number of frames in a dynamic programming template, this difference fades away. What prevents hidden-Markov models from having many states is the difficulty of estimating the resulting large number of transition probabilities.

Hidden-Markov modeling necessarily computes the variability of spectra at different parts of each word. It also has variable time distortion penalties, and it relates these penalties to the spectral distortion penalties in a theoretically defensible way. On the other hand, its timing model is unrealistic in that the probability of staying in a given hidden state decays exponentially with time. The development of more realistic timing models is an active area of research at present [42].

Hidden-Markov modeling has generally proved itself superior to dynamic programming time warping both in laboratory systems and in commercial products (notably those from Verbex [43]). The advantage presumably stems from the inclusion of spectral and timing variability information. On the other hand, their need for multiple examples of each word makes training hidden-Markov systems burdensome. Moreover, it is not yet clear that dynamic programming algorithms with variable spectral and time distortion penalties could not be made to work at least as well.

In the last few years there has been rapidly increasing interest in the application of neural networks to speech recognition, in particular a self-training pattern recognition algorithm called a *multi-layer perceptron* [44]. Such neural networks are even freer of imposed ideas of how speech recognition should operate than the other statistical approaches just discussed. While computationally expensive during the learning phase on serial computers, they are well suited to emerging massively parallel architectures. To my knowledge, no-one has yet described a neural network algorithm for continuous speech recognition, but in isolated-word tests multi-layer perceptrons are beginning to show performance levels comparable with the best alternatives.

Neural networks certainly represent an exciting prospect for speech recognition, but we do need to view them with some caution. In the past, LPC, hidden-Markov modeling and expert systems have all been adopted with frenzied enthusiasm as the technique that would revolutionize speech recognition. Claims are being made for neural network approaches that are just as extreme. There are, for example, assertions that we do not need to worry about the acoustic representation presented to the algorithm because it can work out its own representation from the data it is given. This may not be impossible, but it would probably be exceedingly inefficient. Even if neural networks do model the learning behavior of the brain – which is not established – we should remember that the ear is a hard-wired device, and that, in any case, we may not want to build recognizers that learn language in the same way that babies do. It would be

surprising if neural networks did not make a big contribution to speech recognition; it would be equally surprising if they turned out to be the whole answer.

10. Spectral Representations and Distance Measures

The choice of the way in which speech spectra are represented is closely tied to the way in which the spectra are compared. These two operations are central to the speech recognition process, and they account for most of the computation.

As with vocoders, the first useful spectral representations for speech recognition used filterbanks. Typically, there are between four and twenty channels in the range 0-4 kHz or 0-5 kHz, usually with channel spacing based on a perceptual frequency scale. The energies in each channel are normally expressed on a log scale. In this way, changes in the overall level of the input appear as an additive constant in all the channels, and the spectral profile is preserved. Level changes, which are more likely to signal irrelevant prosodic differences or changes in distance from the microphone than useful phonetic distinctions, can then be easily normalized out.

A common method of measuring the similarity of two spectra is to compute the Euclidean distance over the log channel energies. The Euclidean distance is simply the sum of the squares of the differences in the log energies in the corresponding channels. To avoid multiplications, the Euclidean distance is sometimes replaced by the city-block distance, in which the absolute differences in log channel energies rather than their squared differences are summed.

Since spectral envelopes are smooth, representation by channel energies is inefficient: values in adjacent channels are strongly correlated. A statistical technique called *principle components analysis* [45] provides a transformation that will linearly combine a set of correlated parameters to produce a new, uncorrelated set. The new parameters are ordered in terms of the variance of the data such that a subset of the first n parameters contains the largest proportion of the total variance that can be concentrated into this number of variables. Thus, principle components analysis is an attractive way of describing the variability of spectra in a reduced number of variables. Because the transform involved is *orthonormal* (*i.e.* it corresponds to a rotation in space), Euclidean distances are unaffected provided the full dimensionality of the space is retained.

It turns out that the cosine transform is a close approximation to a principle components analysis of filterbank log energies for speech, and this transform rather than a true principle components representation is commonly used [46]. Since the cosine transform of a log power spectrum is called a *cepstrum,* the cosine transform of the log energies of a mel-scale filterbank is often called a *mel-cepstrum.*

Apart from its useful role in reducing the amount of computation needed in calculating Euclidean distances, taking the cosine transform and retaining only the low-order, high-variance terms can actually improve recognition performance. This is because the low-order terms respond to smooth features in the spectrum, while the higher-order terms are sensitive to spectral fine structure. Harmonics of the fundamental appear as fine structure, and are therefore filtered out by the truncation of the cosine series.

The major alternative to filterbank front-ends for speech recognizers is LPC. A method of comparing spectra known as the *Itakura metric* [47] is computationally attractive because the matrix inversion required for LPC analysis need not be carried out on the frames speech to be recognized: the autocorrelation properties of the speech are simply tested for consistency with the LPC coefficients of the reference data. Increasingly, however, LPC cepstrum coefficients, which can be easily derived from an LPC analysis, are being used. The relative superiority of filterbank and LPC front-ends is still disputed, though it is widely believed that LPC is more sensitive to noise.

Vector quantized representations lead to particularly efficient spectral comparisons because the similarity of any pair of items in the codebook can be stored in a table. Spectral comparison is therefore reduced to table lookup. Vector quantization is particularly well suited to hidden-Markov modeling as we have already noted. However, the quantization does seem to reduce recognizer performance.

There is currently much interest in replacing the simple unweighted calculation of the Euclidean distance over the cepstrum by a distance metric in which the contributions of different cepstrum coefficients are weighted differently. Two separate arguments exist for applying weights, though interestingly they lead to similar values for the weights. The first argument is based on properties of the speech signal and on human speech perception. Formants have a certain range of bandwidths that corresponds to a range of cepstrum components. Cepstrum components below this range are sensitive to phonetically unimportant changes in spectral balance, and components above the range are sensitive to harmonics of the fundamental frequency. Weighting in favor of mid-range cepstrum coefficients therefore enhances sensitivity to formants and provides a spectral similarity measure closer to human judgments of phonetic similarity [48,49].

The second argument for weighting is based on ideas from statistical pattern recognition. Consider a set of samples described in terms of a number of variables – in this case a "sample" would be a spectral frame, and the variables would be channel energies or cepstrum coefficients. If the samples have a multivariate Gaussian distribution about their mean, and if the variables are uncorrelated and have equal variances, then the Euclidean distance of a sample from the mean will be proportional to the log probability of its occurring at this point. Cepstrum coefficients are indeed uncorrelated, and they do have roughly Gaussian distributions, but the variances are certainly not the same for different coefficients. To make the variances look the same, the Euclidean distance contribution of each cepstrum coefficient has to be scaled by its variance [50]. This statistical approach to distance weighting is attractive because it can be applied to representations other than filterbanks, for which the principle components do not look like cosine functions [51].

Both arguments advocate similar steadily increasing weights over roughly the first eight coefficients. Beyond this point, however, the recommendations diverge, in that the statistical method says that the weights should continue to increase, since the variances continue to decrease, while the perceptual argument and practical experience say they should decrease. At the risk of getting too

technical here, we believe that the problem in the statistical method is that the *within-class* variances rather than the total variances should be used [51,52], and that uncertainties in estimating the means should be taken into account. We are actively studying these problems.

We saw in the section on speech analysis and coding that conventional LPC does not reflect the non-uniform frequency resolution of the ear. However, by applying LPC to the output of a simulated mel-scale filterbank a perceptually motivated LPC analysis can be generated. The analysis now no longer corresponds to a production model consisting of an all-pole filter, and it must instead be seen entirely as a spectrum-fitting technique. Encouraging recognition results have been reported, especially when combined with weighted cepstrum distances measures and when tested with noisy data [11].

11. Auditory Models

Humans are more effective at speech recognition than the best automatic systems. Their advantage persists even when they are prevented from using their knowledge of the language by being asked to recognize nonsense words. This suggests that the acoustic representation provided by the ear may be better than those currently used in recognizers. We have already noted that speech has probably evolved to suit the representation provided by the ear, so the advantage is not surprising. Since the superiority of human listeners increases when the speech signal is degraded by noise or distortions, it seems that the representation we use must also be particularly robust. These considerations provide the motivation for using auditory models as front-ends for speech recognizers.

Auditory models present a moving target: as auditory features (such as the use of a perceptual frequency scale) become common, representations need additional auditory properties to be called auditory models. Moreover, some developers of auditory models are aiming to replicate physiological mechanisms [53,54], while others attempt to reproduce psychoacoustic properties. Our view is that blind replication of mechanisms is a mistake. The mechanisms employed by the ear may reflect physiological constraints that are quite different from those encountered in digital implementations. Piecemeal introduction of physiological features (such as half-wave rectification in inner hair cells [55]) may degrade performance when introduced into a necessarily incomplete model [56]. The final test of an auditory model is its contribution to a speech recognition system; but since this criterion provides few pointers to the developer, a useful intermediate criterion is its success in reproducing psychoacoustic properties. Useful psychoacoustic properties include frequency resolution, phase sensitivity, masking phenomena, and details of the response to formants.

Since our auditory model [51] has shown a particularly large recognition performance advantage, it seems worthwhile to describe its structure and properties briefly. The model has 32 channels equally spaced on a perceptual frequency scale with centre frequencies ranging from 100 Hz to 3.3 kHz. Two distinct outputs are produced in parallel. The first, called the *onset detector*, has temporal masking properties, and it provides a distinction between voiced and voiceless speech and silence as well as responding strongly to onsets. The second and more

important output, called the *periodicity detector,* responds mainly to formant structure and has two-tone suppression properties. It is generated by subtracting the log power output of an inhibitory filter from that of an excitatory filter. The result is that each channel consists of a central excitatory region extending midway to the centres of the adjacent channels and surrounded by an inhibitory region. Spectral components in the excitatory region increase the channel output, and components in the inhibitory region reduce it. This process is similar to what is believed to occur in color perception. By carefully designing the filter shapes, published two-tone suppression curves can be matched.

The two-tone suppression properties of the periodicity detector result in some desirable behavior in response to formants. The fourth and fifth formants are largely masked out in voiced speech. Pairs of formants fuse into a single peak when they come within a certain frequency difference of each other, just as they are found to do in perception experiments. The response to the first formant is largely unaffected by the locations of harmonics of the fundamental frequency, a property that had been claimed to be impossible to achieve in an acoustic analysis. Also, the response to phase differences in the model seems close to that shown by the ear.

Both the onset detector and the periodicity detector have a substantial degree of independence of changes in overall level or spectrum balance in the input signal.

A spectrum comparison metric has been developed that combines the two 32-channel outputs into a set of eight numbers. The metric is derived from estimates of the principle components of the within-class covariance matrix. Using this metric, the performance of the model has been compared with that of a conventional mel-cepstrum representation and unweighted Euclidean distance using identical data and identical recognition algorithms. In speaker dependent and independent connected and isolated word tests the model gives reductions in error rates of between two and three. When the test data is degraded by noise or changes in spectrum balance, the advantage shown by the model rises to factors of between six and twenty.

12. Noise and Distortions in Speech Recognition

In all practical applications of speech recognition the speech is subject to some degree of degradation. In some environments -- aircraft cockpits, production lines, and over telephone links, for example -- noise is an obvious problem. But even in office environments, background noise can interfere with speech recognition. In addition, room reverberation and distance from the microphone affect the spectral content of the speech. In most cases, the damage to speech recognition performance is caused not directly by the degradation but by changes in the level of the degradation -- changes in noise level, distance from the microphone, and so on.

Noise in the ears of a speaker invariably causes an increase in the loudness of the speech. Speaking louder not only increases the total energy in the speech but it also raises the fundamental frequency and changes the spectral balance to increase the high-frequency content. Some researchers [57] have found this

indirect effect of noise to have a greater effect on recognition performance than the direct effect of the noise entering the microphone.

Where appropriate, problems can be alleviated by using a head-mounted noise-canceling microphone, such as a *Shure SM10*. This reduces interfering noise, room reverberation effects, and variations in the distance from the mouth to the microphone. If the main source of interfering noise is localized, an additional microphone and two-channel adaptive filtering can in principle be used to cancel its effect on the speech, though I know of no instances of this approach being used. Certainly two-channel filtering does not seem to be useful in aircraft, where the noise sources are multiple and diffuse [58]. Some positive results have been reported in aircraft noise using a throat microphone to supplement the usual microphone mounted in front of the mouth [59]. In many situations, however, the microphone arrangement cannot be controlled, and in high-noise conditions a significant level of noise remains whatever microphone arrangement is used.

When the acoustic analysis is a filterbank or simulated filterbank, resistance to noise can be improved by a so-called *noise marking* technique provided the noise is steady [60]. The noise level in each channel is estimated during periods when no speech is present. Then, when frames are being compared, channels in which the energy is not clearly above the noise level are ignored. It is clearly crucial that the noise level should not change between the time when it is estimated and the times when it is used. Moreover, the method cannot be applied to systems with LPC-based and auditory-model front-ends, in which noise and speech interact in a non-linear manner. With conventional filterbank front-ends, its use requires that the channels should be kept separate in the frame comparison process. This precludes the use of cosine and other transforms that reduce computation and potentially improve performance. Noise marking increases the complexity of the frame comparison process. Computational efficiency here is critical because it is the point where the load is highest, and it increases with vocabulary size. The computational cost of the front-end, by contrast, is independent of vocabulary size.

We have already seen that changes in noise level have effects on the speech itself. If the noise level is steady over the long term it is therefore useful to have the speakers be subjected to the noise during the training phase, perhaps by playing the noise over headphones. If the noise level changes, microphone arrangements or noise marking provide no help in counteracting the effects of changes in the voice.

To my knowledge, only auditory models are robust against both the direct effects of additive noise and the indirect effects through voice changes. Moreover, the noise-resistance properties of auditory models are not confined to steady noises, and the noise level does not need to be estimated. Auditory models are computationally more costly as front-ends. But this is off-set by the simple frame comparisons that can be used, and we have seen that it is particularly important to keep the frame comparison process simple. For these reasons, we believe that auditory models are the best prospect for achieving reliable speech recognition in difficult acoustic conditions. A large proportion of all applications of speech

recognition involve such conditions.

13. Specifying Recognizer Performance

It is natural to want to have a number to specify the performance of a recognizer, much as the breaking strength of a cable can be specified in units of force or the speed of a CPU can be specified in MIPS. Unfortunately, there are many difficulties in trying to define a quantity that can be reproducibly measured and that will usefully specify performance.

Two parameters that are often used to cite recognizer performance are vocabulary size and recognition rate. We will deal with vocabulary size first. Quoted vocabulary size may simply state the capacity of the reference speech memory space, and this may be much larger than the recognizer could handle if it is to provide a usable recognition rate. In recognizers with syntax, it is the average number of choices, known as the *branching factor,* rather than the total vocabulary size that is important (or in systems with probabilistic syntaxes, a generalization of the branching factor known as *perplexity*). In connected-word recognizers branching factors may be limited not only by the need for adequate recognition accuracy but also by the speed of the recognizer in attempting to keep up with the speech being input.

Not all vocabularies of the same size present the same degree of difficulty. Pairs of words that have many different features and few common features are easiest to discriminate. A common, particularly difficult vocabulary is the alphabet, in which sets of letters such as *b, c, d, e, g, p, t* and *v* share a common vowel. The difficulty is not so much that the words in such sets differ in only one phoneme, but rather that most of their duration consists of nominally the same vowel. There will inevitably be random variations in this vowel, and these variations will occur in the training data as well as in the test data. Thus, the vowel in a test utterance of *c* may be closer to that in the reference version of *t* than to that in the reference version of *c,* and this spurious similarity can overwhelm the true similarity in the initial part of the *c*'s. With the use of more training examples, the random variations will be averaged out, and in systems using speech-sound rather than whole-word reference units the problem does not arise. This means that the difficulty of a vocabulary depends on the details of the recognition system with which it is used.

Recognition accuracy is sometimes quoted without reference to the vocabulary on which it was obtained. This is clearly absurd. However, even when the vocabulary is specified, there are serious difficulties with the measurement of performance. Performance measurement is a statistical sampling process that one hopes would provide similar estimates in repeated independent tests. If the performance level is high — as it has to be for many applications — errors are rare and estimation of the error rate requires large amounts of test data.

Generally, performance estimates are sought not for a particular speaker or set of speakers, but for the whole population. But performance varies greatly from one speaker to another and even for one speaker from one occasion to another. Performance estimates would need to be averaged over a large speaker set to average out speaker differences.

Recognition performance is often quoted as the percentage of words correctly recognized. This practice may give the misleading impression that an improvement from 75% correct to 90% correct is more significant than an improvement from, say, 95% to 99%. The reduction in the total number of errors is indeed greater in the first case, though the increase in the mean time between errors (which is proportional to the reciprocal of the error rate) is less, and time between errors may be what controls a user's subjective impression of system reliability. From the developer's point of view, the second case represents much more effort than the first: approaching 100% recognition accuracy is like approaching absolute zero in low-temperature physics. A more important consideration is the following: if several systems are tested in different conditions — with different speakers or in quiet and noisy conditions, for example — the relationship that tends to be preserved between them is the ratio of the error rates. Thus, if system A makes twice as many errors as system B in one condition, it is likely to make twice as many errors in another condition. If the two systems score 98% and 99% in the first condition, and in the second condition system A scores 80%, then system B's performance is more likely to fall to 90% than to 81%. This rule of thumb suggests that we should always quote error rates rather than percentages correct. Then in comparing systems, rather than looking at differences in error rates we should consider their ratios — or, equivalently, the differences in the log error rates.

The arguments just presented raise a problem when we try to estimate average recognition results, say across different speakers. Directly averaging error rates or proportions correct would give disproportionate weight to problem speakers. Averaging log error rates — amounting to computing the geometric mean of the error rates — would seem to be a better procedure. However, the influence of "noise" (*i.e.* inaccuracies in the estimates) on the log error rates is much greater for low error rates: a moderate-sized test of an effective system may result in no errors, though the true error rate can never be zero and the log of zero is unbounded. Averaging a set of log error rates containing one case with no errors would always give an apparent average error rate of zero. There may be a sophisticated statistical technique for handling this problem, but there seems to be no simple, optimal method of averaging recognition results.

A method of evaluating performance has been proposed [61] that would potentially make more effective use of a performance test by using measures of similarity of word patterns derived by a recognizer rather than just using the one-bit right-or-wrong recognition decision. The method could also potentially eliminate the vocabulary dependence of performance measurement. Match distances are recorded between pairs of reference templates and between individual words to be recognized and their templates. Multidimensional scaling is then used to determine a placement of the words in a space that is consistent with the distance data. Templates are imagined to correspond to the points centred on the spherical distributions of the individual examples. The performance measure, the *equivalent vocabulary capacity* or EVC, is the number of such spheres that can be packed into the space with a given amount of overlap. The method is in principle independent of the vocabulary used provided that the vocabulary is diverse enough to sample the space adequately. Similarly, the difficulty of a specific

vocabulary, called the *required vocabulary capacity* or RVC, can be assessed by computing the EVC that would be necessary to cope with it. These ideas offer a potentially powerful method of performance evaluation, but they have still to be fully validated experimentally.

Disturbingly, when two laboratories test the same system with the same database they do not always obtain the same result. Gain settings may be different, for example. When the NATO Research Study Group on Speech Processing, RSG10, organized the testing of one commercial recognizer at two sites, dramatically different results were obtained. It emerged that one of the testing sites had delegated the task to some enthusiastic but inexperienced staff, who, on obtaining an error, repeatedly re-input the word at different gain settings until in was correctly recognized or they were satisfied that it was impossible. In this way they managed to reduce the error rate to a fraction of that found by the other group, and if theirs had been the only result reported it would never have been questioned. Laboratories tend to obtain better results with their own systems than with others because they know how to set up their own systems for best performance. A widely reported test from one internationally respected group found their own algorithm to work much better than an alternative proposed by another group. The alternative algorithm contained a parameter whose value had been set far from its optimal value. When a third group repeated the test with an optimized value, they found a slight advantage for the alternative.

At this point it is worth considering the motivations for performance estimation. An investigator may wish to predict the *absolute* performance of a recognition system in a particular application, or he may wish to determine the *relative* performance of two or more systems. The former aim is the more difficult to satisfy: workers at Verbex reported that error rates might increase by an order magnitude between tests in the laboratory and tests in the field. Performance in a real task also changes over time as users gain experience.

When the aim is rather to rank the performances of systems, some of the sources of variation in estimates can be eliminated by using exactly the same material to test the systems. Standard databases are being developed, and the U.S. National Bureau of Standards is particularly active in this effort [62]. Gillick has recently pointed out [63] that the sensitivity of comparisons can be improved further by noting whether systems being compared make errors on the same words rather than simply comparing overall error rates.

One danger with using standard databases is that systems under development will be tested repeatedly, and consciously or unconsciously they will be adapted to fit the peculiarities of the database. This problem of adaptation to the test data is common in many areas of pattern recognition.

Overall error rates do not, in any case, tell the whole story about performance. First, it is important that recognizers should reject words not in the vocabulary: users often inadvertently present such words to recognizers. Test databases free of spurious items do not test this ability; and even if they did, the frequency and type of these items would have to be fixed arbitrarily. Recognizers often have adjustable rejection thresholds that allow probability of accepting spurious inputs to be traded off against probability of rejecting valid inputs.

In addition, overall error rates say nothing about the way in which errors are distributed across the vocabulary. To take an extreme example, 91% recognition accuracy in a digit recognition task might mean 91% accuracy on each digit or 100% accuracy on the digits 0 to 8 and a random 10% on the digit 9. The latter case offers more potential for improvement by changing the vocabulary or modifying the algorithm. However, if neither the vocabulary nor the algorithm can be changed, the even distribution of errors is preferable. To see this, consider the task of having the recognizer correctly recognize each of the digits once: the even distribution case would on average require eleven digits to be input, while the uneven case would on average require nineteen (ten repetitions of the digit 9).

Connected-word recognizers present particular difficulties for performance evaluation because of the possibility of insertions and deletions as well as substitution errors. If, for example, the digit string "990" is presented and the system responds with "9950," it is not clear whether a 5 is inserted before the 0 or the second 9 was misrecognized as a 5 and an additional 9 is inserted.

One can regard any word sequence containing an error as entirely wrong; but this is an inefficient use of the data and it says nothing about the kinds of errors being made. Another strict analysis compares the input and output in order: in the example above, the 5 in the output would be compared with the 0 in the input and the 0 in the output would be taken to be an insertion. This analysis is pessimistic and insensitive to true error rates unless insertions and deletions are known to be rare.

A popular evaluation strategy uses dynamic programming to match a sequence of symbols representing the words found by the recognizer to a sequence representing the words actually presented to the recognizer. Given penalty costs for insertions, deletions and substitutions, the method finds the lowest-cost — hence the most *favorable* — interpretation of the recognizer output in terms of the input. This will generally underestimate the number of errors made by the recognizer, but simulations have shown that for error rates below about 20% it is reasonably accurate [64].

Finally, if the word boundaries found by the recognizer can be determined, and if the true word boundaries in the test data are known, a method using dynamic programming to compare boundary locations can give accurate performance estimates over the full performance range [64]. Since the symbol-matching approach works well at practical performance levels, the main return for the additional effort required by the boundary-matching approach is in the diagnostic information it provides. Such information is of more interest to algorithm developers than to applications-oriented workers.

14. Concluding Remarks

We can all speak, and we can all understand speech. Since most of the processing involved in both of these actions is automatic and unconscious, it is hard to realise what remarkable skills they are. We are led to underestimate the difficulties of conferring similar skills on machines; and those who want to sell speech I/O products often reinforce the misimpression by understating the

difficulties and exaggerating the performance of their product. The message of much of this report has therefore been rather negative: not only are speech input and output difficult tasks, it is also difficult even to measure the performance of a speech input or output system. In case the balance has tilted too far the other way, we should hasten to remember that useful things can be done with technology available today, and progress — though slower than some accounts would claim — is being made.

Communication between humans is overwhelmingly by voice. Yet communication with machines is at present even more overwhelmingly by touch and sight. Speech input and output is generally more natural and often faster than using keyboards and screens. It also leaves hands and eyes free for other tasks, and allows remote access with equipment no more complex than a wireless microphone or a telephone handset. Given these advantages, it seems certain that as speech technology progresses it will play an increasingly important role in our interaction with machines.

Readers wanting a more comprehensive introduction to speech technology might try the book by Parsons [65].

Acknowledgments

The production of this report was stimulated and partially funded by Applied AI Systems, Inc., as part of a contract from Transport Canada. The Department of National Defence through the Defence and Civil Institute of Environmental Medicine (DCIEM) has provided long-term financial support for the NAE Speech Research Centre.

I am grateful to Claude Lefèbvre for help in preparing the references.

REFERENCES

1. M.J. Hunt, "The Speech Signal," *Proc. NATO AGARD Lecture Series No. 129, Speech Processing*, pp. 2.1-2.12, 1983.
2. J.D. O'Connor, *Phonetics*, Penguins Books, Harmondsworth, Middlesex, England, 1973.
3. M.J. Hunt, "Speech is More Than Just an Audible Version of Text," to appear in *Structure of Multimodal Dialogues including Voice*, eds M.M. Taylor and F. Néel, North Holland, probably 1988.
4. M.J. Hunt, "Studies of Glottal Excitation using Inverse Filtering and an Electroglottograph," *Proc. XI'th Intl. Congress of Phonetic Sciences*, Tallinn, Estonia, August 1-7, 1987, Vol. 3, pp. 23-26.
5. J.O. Pickles, *An Introduction to the Physiology of Hearing*, Academic Press, London, 1982.
6. J.V. Tobias, *Foundations of Modern Auditory Theory*, Academic Press, New York, 1970, 2 Volumes.
7. C.G.M. Fant, "Acoustic Description and Classification of Phonetics Units," in *Speech Sounds and Features*, MIT Press, Cambridge, 1973.
8. E. Zwicker & E. Terhardt, "Analytical Expressions for Critical-Band Rate and Critical Bandwidth as a Function of Frequency," *J. Acoust. Soc. America*, Nov. 1980, Vol. 68, pp. 1523-1525.
9. D.H. Klatt, "Prediction of Perceived Distance from Critical Band Spectra. A First Step," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-82*, Paris, May 1982, pp. 1278-1281.
10. L.C.W. Pols, "How to Make Efficient Use of the Fact that the Speech Signal is Dynamic and Redundant," *Proc. IEEE Int Conf. on Acoustics, Speech and Signal Processing, ICASSP-86*, Tokyo, April 1986, pp. 1963-1966.
11. H. Hermanski, "An Efficient Automatic Speaker-Independent Speech Recognition by Simulation of some Properties of Human Auditory Perception," *Proc. IEEE Int Conf. on Acoustics, Speech and Signal Processing, ICASSP-87*, Dallas, April 1987, pp. 1159-1162.
12. J.N. Holmes, "The JSRU Channel Vocoder," *IEE Proc. Communications*, Feb. 1980, Vol. 127, part F, pp. 53-60.
13. J.D Markel & A.H. Gray, *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976.
14. C.K. Un & D.T. Magill, "The Residual-Excited Linear Prediction Vocoder with Transmission Rate below 9.6 Kbits/s," *IEEE Trans. on Comm.*, Dec. 1975, Vol. COM-23, pp. 1466-1474.
15. B.S. Atal & J.R. Remde, "A new Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," *Proc. IEEE Int Conf. on Acoustics, Speech and Signal Processing, ICASSP-82*, Paris, May 1982, pp. 614-617.

16. A.K. Krishnamurthy & D.G. Childers, "Two-channel Speech Analysis," *IEEE Trans. on Acoustics, Speech and Signal Processing,* 1986, Vol 34, pp. 730-743.

17. M.J. Hunt & C.E. Harvenberg, "Generation of Controlled Speech Stimuli by Pitch-Synchronous LPC Analysis of Natural Utterances," *Proc. Int. Congress on Acoustics,* Toronto, July 1986, Vol. 1, paper A4-2

18. A.J. Fourcin & E. Abberton, "First Applications of a new Laryngograph," *Medical and Biological Illustration,* Vol. 21, pp. 172-182.

19. R.M. Gray, "Vector Quantization," *IEEE ASSP Magazine,* April 1986, Vol. 1, no. 2, pp. 4-29.

20. J. Allen, S. Hunnicutt & D. H. Klatt, *From Text to Speech: The MITalk System,* Cambridge University Press, Cambridge, England, 1987.

21. J.N. Holmes, I.G. Mattingley & J.N. Shearme, "Speech Synthesis by Rule," *Lang. & Speech,* 1964, Vol. 7, pp. 127-143.

22. M.M. Taylor & M.J. Hunt, "Flexibility versus Formality," to appear in *Structure of Multimodal Dialogues including Voice,* eds M.M. Taylor and F. Néel, North Holland, probably 1988.

23. D.B. Pisoni, "Perceptual Evaluation of Voice Response Systems," *Proc. Workshop on Standardization for Speech I/O Technology,* Gaithersburg, MD, March 1982, pp. 185-192.

24. J.P. Olive, "Rule Synthesis of Speech from Dyadic Units," *Proc IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-77,* New York, May 1977, pp. 568-570.

25. O. Fujimura, M.J. Macchi & J.B. Lovins, "Demisyllables and Affixes for Speech Synthesis," *Proc. 9th Int. Congr. on Acoustics,* Madrid, 1977, paper I107.

26. Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J. Makoul, P.J. Price, S. Roucos & M. Schwartz, "Byblos: The BBN Continuous Speech Recognition System," *Proc. DARPA Speech Recognition Workshop,* San Diego, March 1987, pp. 1-5.

27. A.I. Rudnicky, Z. Li & L.K. Baumeister, "The Lexical Access Component of the CMU Continuous Speech Recognition System," *Proc. DARPA Speech Recognition Workshop,* San Diego, March 1987, pp. 18-21.

28. C. Weinstein, MIT Lincoln Laboratory, personal communication, Fall 1987.

29. M.J. Hunt, "Delayed Decisions in Speech Recognition - the Case of Formants," *Pattern Recognition Letters,* 1987, Vol 6, pp. 121-137.

30. V.W. Zue, "Experiments on Spectrogram Reading," *Proc IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-79,* Washington DC, April, 1979, pp. 116-119.

31. J.S. Bridle, R.M. Chamberlain & M.D. Brown, "An Algorithm for Connected Word Recognition," *Proc IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-82,* Paris, May 3-5, 1982, pp. 899-902.

32. L.R. Rabiner, S.E. Levinson, A.E. Rosenburg & J. Wilpon, "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques," *IEEE Trans. on Acoustics, Speech and Signal Processing,* 1979, Vol ASSP-27, pp. 336-349.

33. M.J. Hunt, "Speaker Adaptation for Word-Based Speech Recognition Systems," Spring 81 Meeting, Acoust. Soc. America, Ottawa, abstract published in, *J. Acoust. Soc. America,* Vol. 69, pp. S41-42, 1981.

34. M.J. Hunt & P. Mermelstein, "Normalization of Speech for Automatic Recognition," *Final Report, Govt. of Canada DSS Contract No. 8SR79-00048,*

35. K. Choukri, G. Chollet & Y. Grenier, "Spectral Transformations Through Canonical Correlation Analysis for Speaker Adaptation in ASR," *Proc IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-86,* Tokyo, April 1986, pp. 2659-2662.

36. A.E. Rosenburg, L.R. Rabiner, S.E. Levinson & J. Wilpon, "A Preliminary Study on the Use of Demisyllables in Automatic Speech Recognition," *Proc IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-81,* Atlanta, April 1981, pp. 967-970.

37. M.J. Hunt, M. Lennig & P. Mermelstein, "Experiments in Syllable-based Recognition of Continuous Speech," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-80,* Denver, April 1980, pp. 880-883.

38. D. Sankoff & J.B. Kruskal, *Time Warps, String Edits and Macromolecules: Theory and Practice of Sequence Comparison,* eds Sankoff and Kruskal, Addison Wesley, Reading Mass, 1983.

39. L.R Bahl, F. Jelinek & R.L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* March 1983, Vol. PAMI-5, pp. 179-190.

40. H. Bourlard, Y. Kamp, H. Ney & C.J. Wellekens, "Speaker-Dependent Connected Speech Recognition via Dynamic Programming and Statistical Methods," in *Speech and Speaker Recognition,* ed. M.R. Schroeder, Karger, Basel, 1985, pp. 115-148.

41. S.E. Levinson, L.R. Rabiner & M.M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *Bell System Tech. J.,* 1983, Vol. 62, pp. 1035-1074.

42. M.J. Russell & A. E. Cook, "Experimental Evaluation of Duration Modelling Techniques for Automatic Speech Recognition," *Proc IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-87,* Dallas, April 1987, pp. 2376-2379.

43. S.L. Moshier, "Talker Independent Speech Recognition in Commercial Environments," *J. Acoust. Soc. America,* 1979, Vol. 65, pp. S132.

44. A. Waibel, T. Hanazawa, K. Shikano, G. Hinton & K. Lang, "Phoneme Recognition: Neural Networks vs. Hidden Markov Models," *Proc IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-88,* New York, April 1988, paper 8.S.3.2.

45. L.C.W. Pols, L.J.Th. van der Kamp & R. Plomp, "Perceptual and Physical Space of Vowel Sounds," *J. Acoust. Soc. America,* Nov. 1980, Vol. 46, pp. 458-467.

46. S.B. Davis & P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustics, Speech and Signal Processing,* Aug. 1980, Vol. ASSP-28, pp. 357-366.

47. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing,* Feb. 1975, Vol. ASSP-23, pp. 67-72.

48. F. Itakura & T. Umezaki, "Distance Measure for Speech Recognition Based on the Smoothed Group Delay Spectrum," *Proc IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-86,* Tokyo, April 1986, pp. 1257-1260.

49. K. Paliwal, "On the Performance of the Quefrency-Weighted Cepstral Coefficients in Vowel Recognition," *Speech Communication,* 1982, Vol. 1, pp. 151-154.

50. Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *Proc. IEEE Int Conf. on Acoustics, Speech and Signal Processing, ICASSP-86,* Tokyo, April 1986, pp. 761-764.

51. M.J. Hunt & C. Lefèbvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-88,* New York, April 1988, paper 15.S.5.9.

52. M.J. Hunt, "A Statistical Approach to Metrics for Word and Syllable Recognition," Fall 79 Meeting, Acoust. Soc. America, Salt Lake City, abstract published in *J. Acoust. Soc. America,* 1979, Vol 66, pp. S535-536.

53. R.F. Lyon, "A Computational Model of Filtering, Detection, and Compression in the Cochlea," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-82,* Paris, April 1982, pp. 1282-1285.

54. S. Seneff, "Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-84,* San Diego, March 1984, paper 36.2.1.

55. M.J. Hunt & C. Lefèbvre, "Speech Recognition Using an Auditory Model with Pitch-Synchronous Analysis," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-87,* Dallas, April 1987, pp. 813-816.

56. M. Blomberg, R. Carlson, K. Elenius & B. Granstrom, "Experiments with Auditory Models in Speech Recognition," in *The Representation of Speech in the Peripheral Auditory System,* eds. Carlson and Granstrom, Elsevier Biomedical Press, Amsterdam, 1982, pp. 197-201.

57. P.K. Rajesekran, G.R. Doddington & J.W. Picone, "Recognition of Speech under Stress and in Noise," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-86,* Tokyo, April 1986, pp. 733-740.

58. G.A. Powell, P. Darlington & P.D. Wheeler, "Practical Adaptive Noise Reduction in the Aircraft Cockpit Environment," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-87,* Dallas, April 1987, pp. 173-176.

59. V. Viswanathan, C. Henry, R. Schwartz & S. Roucos, "Evaluation of Multisensor Speech Input for Speech Recognition in High Ambient Noise," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-86,* Tokyo, April 1986, pp. 85-88.

60. J.N. Holmes & N.C. Sedgwick, "Noise Compensation for Speech Recognition using Probabilistic Models," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-86,* Tokyo, April 1986, pp. 741-744.

61. M.M. Taylor, "Issues in the Evaluation of Speech Recognition Systems," *J. American Voice I/O Soc.,* 1986, Vol. 3, pp. 39-67.

62. D.S. Pallett, "Test Procedures for the March 1987 DARPA Benchmark Tests," *Proc. DARPA Speech Recognition Workshop,* San Diego, March 1987, pp. 75-78.

63. L. Gillick, oral presentation at DARPA Speech Recognition Meeting, Cambridge, Mass, October 1987.

64. M.J. Hunt, "Evaluating the Performance of Connected-Word Speech Recognition Systems," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-88,* New York NY, April 1988, paper 34.S.10.13.

65. T.W. Parsons, *Voice and Speech Processing,* McGraw-Hill, New York, 1986.

REPORT DOCUMENTATION PAGE / PAGE DE DOCUMENTATION DE RAPPORT

REPORT/RAPPORT 1a NAE-AN-50	REPORT/RAPPORT 1b NRC No. 28714
REPORT SECURITY CLASSIFICATION CLASSIFICATION DE SÉCURITÉ DE RAPPORT 2 UNCLASSIFIED	DISTRIBUTION (LIMITATIONS) 3 UNLIMITED

TITLE/SUBTITLE/TITRE/SOUS-TITRE
4 An Overview of Technology for Spoken Interaction with Machines

AUTHOR(S)/AUTEUR(S)
5 M.J. Hunt

SERIES/SÉRIE
6 Aeronautical Note

CORPORATE AUTHOR/PERFORMING AGENCY/AUTEUR D'ENTREPRISE/AGENCE D'EXÉCUTION
7

SPONSORING AGENCY/AGENCE DE SUBVENTION
8 National Research Council Canada
National Aeronautical Establishment Flight Research Laboratory

DATE	FILE/DOSSIER	LAB. ORDER COMMANDE DU LAB.	PAGES	FIGS/DIAGRAMMES
9 Feb. 1988	10	11	12a 36	12b

NOTES
13

DESCRIPTORS (KEY WORDS)/MOTS-CLÉS
14
1. Speech recognition
2. Speech — Technology

SUMMARY/SOMMAIRE
15

This report provides a non-mathematical introduction to speech input and output technology. It is divided into three parts. The first presents necessary background information on speech: on its nature, its production and perception, and on methods of analysis and coding used in speech I/O. A central message is that our subjective impression of speech is misleading and causes us to underestimate the complexity of speech communication. The second part is concerned with speech output and discusses the trade-offs that must be made between the quality and flexibility of the speech generated and the complexity and storage requirements of the speech output system. The final — and longest — part of the report deals with speech recognition. Arguments are presented in favor of statistical rather than rule-based approaches to speech recognition. The categories of recognizer currently available and the algorithms they use are briefly described, with the general conclusion that the performance obtained depends critically on the training process: on the type and quantity of the training material and on the amount of information derived from it. Three more detailed sections cover spectral representations and distance measures, the particular set of representations classed as auditory models, and techniques for handling noise and distortions. The last section discusses the difficulties of specifying recognizer performance, and recommends that all performance measurements should be treated with circumspection.